Multiple-breed genomic evaluation by principal component analysis in small size populations

Gaspa, G.; Jorjani, H.; Dimauro, C.; Cellesi, M.; Ajmone-Marsan, P.; Stella, A.; Macciotta, N. P. P.,

Multiple-breed genomic evaluation by principal component analysis in small size populations,

Animal, 2015, 9 (5), 738–749,
DOI: 10.1017/S1751731114002973.

Multiple-breed genomic evaluation by principal component analysis in small size populations

G. Gaspa[1], H. Jorjani[2], C. Dimauro[1], M. Cellesi[1], P. Ajmone-Marsan[3], A. Stella[4] and N. P. P. Macciotta[1].

[1] *Dipartimento di Agraria, Viale Italia 39, 07100 Sassari, Italy*

[2] *Interbull Centre, Box 7023, S-75007 Uppsala, Sweden*

[3] *Istituto di Zootecnica, Università Cattolica del Sacro Cuore, Piacenza, 29100, Italy*

[4] *Parco Tecnologico Padano, 26900 Lodi, Italy*

Corresponding author: Giustino Gaspa. Email: gigaspa@uniss.it

Running head: Multiple Breed Genomic Selection

Abstract

In this study, the effect of breed composition and predictor dimensionality on the accuracy of direct genomic values in a multi-breed cattle population was investigated. A total of 3559 bulls of three breeds were genotyped at 54001 Single Nucleotide Polymorphisms: 2093 Holstein (H), 749 Brown Swiss (B) and 717 Simmental (S). Direct genomic values (DGV) were calculated using a Principal Component approach for either single (SB) or multiple breed (MB) scenarios. Moreover, DGV were computed using all SNP genotypes simultaneously with SNPBLUP model as comparison. Seven datasets were used: three with a single breed each, three with different pairs of breeds (HB, HS and BS), and one with all the three breeds together (HBS), respectively. Editing was performed separately for each scenario. Reference populations differed in breed composition, whereas the validation bulls were the same for all scenarios. The number of SNPs retained after data editing ranged from 36521 to 41360. Principal components (PC) were extracted from actual genotypes. The total number of retained PC ranged from 4029 to 7284 in Brown Swiss and HBS respectively, reducing the number of predictors by about 85% (from 82% to 89%). Three traits were considered: milk, fat, and protein yield. Correlations between deregressed proofs and direct genomic values were used to assess prediction accuracy in validation animals. In the SB scenarios, average DGV accuracy did not substantially change when either SNPBLUP or PC were used. Improvement of DGV accuracy were observed for some traits in Brown Swiss, only when MB reference populations and PC approach were used instead of SB-SNPBLUP (+10% HBS, +16%HB for milk yield and +3% HBS and +7% HB for protein yield, respectively). With the exclusion of the abovementioned cases, similar accuracies were observed using MB reference population, under the PC or SNPBLUP models. Random variation due to sampling effect or size and

composition of the reference population may explain the difficulty in finding a defined pattern in the results.

Keywords: genomic selection, reference population, multi-breed, dairy cattle, small population.

Implication

A multiple breed approach for predicting direct genomic values in three cattle breeds is presented. The use of multiple breed reference populations might help to increase genomic selection accuracy in small cattle populations. This approach is extendable to populations of other species with reduced number of genotyped animals.

Introduction

Dense marker maps are currently used in the dairy cattle industry for predicting genomic enhanced breeding values (GEBV) in genomic selection (GS) programs (Meuwissen *et al.*, 2001). The advantages of GS in cattle have been extensively reviewed (Hayes *et al.*, 2009a, VanRaden *et al.*, 2009). GEBV accuracy is related to the size and structure of the reference population, the level of linkage disequilibrium (LD) between markers and QTL, the number of QTL underlying the trait and its heritability. Among them, the size of the reference population probably plays the key role to accomplish the theoretical expectations of GS (Goddard and Haye s, 2009).

The need for increasing the size of the reference population for improving GEBV accuracy led to the creation of consortia among breed associations and breeding companies. Thus, genotypes have been exchanged and larger common reference populations have been created as, for instance, in Holstein (Lund *et al.*, 2011) and Brown Swiss (Jorjani *et al.*, 2012). The problem still remains in small or admixed populations. Some authors proposed to use prediction equations estimated in a breed with a large reference population for calculating GEBV in others of small size. Poor results have been obtained, especially for populations that are genetically distant (Hayes *et al.*, 2009b, Pryce *et al.*, 2011, Olson *et al.*, 2012). The use of a multi-breed (MB) reference population could be an alternative for improving GEBV accuracy in small populations. The MB rationale relies on the use of statistical models able to capture LD between SNPs and QTLs when different breeds are analyzed jointly. The combination of different breeds in a larger reference population was simulated by de Roos *et al.*, (2009). The authors concluded that a large marker density was needed to preserve the marker-QTL association across breed, when genetically divergent breeds were pooled together. Furthermore, Kizilkaya *et al.*, (2010) reached the same conclusions simulating MB performances from actual 54K

77 genotypes. A slight improvement in the accuracy of genomic predictions was achieved in
78 real data using medium density chip in MB populations. To date, the increase of marker
79 density (e.g. the use of BovineHD beadchip, Illumina inc., CA) hardly improved GEBV
80 accuracy both in pure and multi breed cattle populations (Harris *et al.*, 2011, Erbe *et al.*,
81 2012, VanRaden *et al.*, 2013).

82 Two main approaches have been proposed in the MB framework: SNP effect
83 estimation (GBLUP or Bayesian methods) from a MB reference considered as
84 homogenous population (Hayes *et al.*, 2009b, Brondum *et al.*, 2011, Pryce *et al.*, 2011), or
85 adaptation of multiple-trait model to the MB case. For instance, Makgahlela *et al.*, (2013)
86 proposed a multiple-trait random regression model, fitting breed proportions as random
87 predictors and an interaction between marker and breed effects. Similar approaches have
88 been implemented by Olson *et al.*, (2012) and Karoui *et al.*, (2012) in US and French MB
89 dairy cattle population, respectively. Although these models allow marker effects to differ
90 among breeds, no or slight gain in accuracy were obtained in comparison with less
91 computational intensive models.

92 An interesting option for across breed genomic evaluation may be represented by
93 the use of multivariate statistics. Principal component analysis (PCA) originally proposed
94 to take into account population structure in human genetics by Cavalli-Sforza (Patterson *et*
95 *al.*, 2006), is currently used in animal breeding for several purposes. In the GS framework,
96 PCA has been used to reduce the number of predictors in the estimation of Direct
97 Genomic Values (DGV) by Solberg *et al.*, (2009). Furthermore, eigenvalues of SNP
98 correlation matrix were also used as variance priors to estimate DGV in simulated and real
99 cattle data (Macciotta *et al.*, 2010, Pintus *et al.*, 2012). In this context, PCA was used to
100 reduce the computational demand and the co-linearity among predictors to calculate DGV

of pure breed animals. Daetwyler *et al.*, (2012) developed a PCA approach to correct for population structure in a complex MB sheep population.

The overall objective of this work was to test the effect of the use of principal components instead of SNP genotypes as predictors in the calculation of direct genomic values either in single (SB) or multi breed scenarios. In particular, the effects of the size and the composition of the multi breed reference population on DGV accuracy were investigated.

Materials And Methods

Data

A total of 3559 bulls of three Italian breeds (2093 Italian Holstein, 749 Italian Brown Swiss and 717 Italian Simmental) were genotyped at 54K SNP. Animals were genotyped with both Illumina Bead chip v1 and v2 that hold 54001 and 54069 SNPs, respectively. Therefore, only common markers (52340) were retained. Seven scenarios of breed composition were considered: Holstein, Brown Swiss, Simmental, Holstein+Brown Swiss+Simmental (HBS), Holstein+Brown Swiss (HB), Brown Swiss+Simmental (BS) and Holstein+Simmental (HS), respectively (Table 1). Bulls with poor quality genotypes (call rate <97.5%) were discarded. Furthermore, checks for Mendelian inconsistency were performed within each breed examining sire-son pairs (animal with >2% inconsistency were eliminated). Finally, bulls with missing phenotypic records were included in the dataset to perform PCA but excluded from the DGV estimation.

Table 1

Quality control was performed separately in each data set. The causes of SNP elimination are summarized in Table 2. SNP with minor allele frequency (MAF) lower than 5% were discarded (monomorphic SNP ranged from 8% to 12% of the total number of SNP). SNP with callrate <97.5% (approximately 3% of total) were eliminated. SNP out of Hardy-Weinberg (Bonferroni corrected P<0.01) were removed in SB scenarios. SNP that deviated from the HW equilibrium in the MB scenarios (HBS, HB, HS and BS) were retained in order to preserve markers potentially able to discriminate among breeds. Moreover, a high percentage of SNP would have not passed this test in a mixed population. The number of SNP retained after data editing ranged from 36521 (Brown Swiss) to 39240 (Holstein) in the case of single breed and from 39615 (BS) to 41360 (HBS) across breed, respectively (Table 2). For the remaining missing values (<0.5% of the total), alleles were imputed using the most frequent allele at each involved locus within each breed.

Table 2

Animals born before December 31st 2000 were included in the reference whereas those born >2000 represented the validation either in SB or MB scenarios. Within each MB scenario the reference populations were set up pooling together bulls belonging to different breeds according to the date of birth. The validation population included always the same bulls across different scenarios (634 Holstein, 171 Simmental and 141 Brown Swiss). Phenotypes used were deregressed proofs (DRGP) provided by the 3 breed associations and calculated separately for each breed. Procedure of Interbull's deregression were carried out in order to remove the effect of pedigree. Moreover, phenotypes of sires that had daughters in foreign countries were corrected according to the multiple across country evaluation (MACE) EBVs for Simmental and Brown Swiss. For Holstein a set of effective daughter contributions (EDC) consistent with the set of

150 reliabilities and the pedigree was derived iteratively. Then full animal model deregression
151 was performed using those EDCs by iteratively finding a set of DRGP consistent with the
152 set of proofs. This procedure is similar to Interbull's deregression, with two differences,
153 namely lack of genetic groups and treating MACE proofs on the Italian scale as if they
154 were domestic proofs (Biffani, Personal communication). In order to have SNP effects
155 comparable across breeds, DRGP (within and across breeds) were standardized to mean
156 = 0 and s.d. = 1. Three traits were considered: Milk Yield (MY), Fat Yield (FY) and Protein
157 Yield (PY). Average DRGP reliabilities for yield traits were 0.93±0.02 (0.90±0.04),
158 0.90±0.07 (0.81±0.06) and 0.88±0.06 (0.85±0.05) in Holstein, Brown Swiss and Simmental
159 reference (validation) bulls, respectively.
160
161 *Principal Component Analysis*
162 The genotype at each locus was coded as -1 and 1 for the opposite homozygotes and 0
163 for the heterozygotes, respectively. PCA was carried out by chromosome in the whole
164 population (reference+validation). PC scores were computed separately for each
165 chromosome in the different scenarios (SB or MB) (Pintus *et al.*, 2012). This chromosome-
166 wise approach was aimed at handling, whenever possible, full rank correlation matrices.
167 The rank of a matrix is defined as the maximum number of independent rows (or columns).
168 For SNP genotype data matrix, the rank is lower or equal to the minimum value between
169 number of animals and number of SNP. In case of small reference population size, the
170 number of observations << number of SNP. Thus the marker (co)variance matrix is not full
171 rank, resulting in a reduction of the maximum number of PC that can be potentially
172 extracted. Previous results obtained on simulated data showed no differences in DGV
173 accuracies between chromosome-wide or genome-wide PC extraction (Macciotta *et al.*
174 2010). Differently from the abovementioned papers, where the number of PC was chosen

based on the proportion of variance explained, in the present investigation the MINEIGEN criterion was adopted (Kaiser, 1960). In particular, for each chromosome a principal component was retained if its eigenvalue was greater than the average (i.e. one in the case PC are extracted from correlation matrices). Finally, individual PC scores were calculated combining the eigenvectors of correlation matrices and original genotypes.

Genomic selection models

Genomic predictions were obtained within each breed using either all marker genotypes available (SB-SNPBLUP) or PC scores (SB-PC) as predictors. The SB-SNPBLUP was considered as the base scenario for comparison with the other approaches. DGV for the different MB sets also were calculated using either SNP genotypes (MB-SNPBLUP) or PC scores (MB-PC).

SB-SNPBLUP. Effects of the SNP were estimated using the following model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{e} \qquad [1]$$

where **y** is a vector of DRGP standardized across breeds with mean 0 and $\sigma_y^2 = 1$, **1** is a vector of ones, μ is the general mean, **Z** is the matrix of SNP genotypes coded as -1, 0 and 1, **g** is a vector of random SNP effects **g** ~ N(0, $\mathbf{I}_m \sigma_g^2$) and **e** is a vector of random residuals **e** ~ N(0, $\mathbf{I}_n \sigma_e^2$), where *m* and *n* are the number of markers and the number of animals, respectively. Variance components $\hat{\sigma}_e^2$ and $\hat{\sigma}_g^2$ and SNP effects were estimated running a Gibbs sampling using 100000 cycles and thinning interval of 10 (20000 samples were discarded as burn in). Estimated variance components were successively used to run a SNP-BLUP model. GS3 software was used to perform the analysis (Legarra *et al.*, 2012).

199

200 *SB-PC.* The effects of PC scores on phenotypes were estimated with model [1] by
201 replacing genotypes with PC scores in **Z**. For *j*-th breed, mixed model equations were set
202 up using as lambda $\lambda^{j} = \sigma^2_{ej} \Big/ \frac{\sigma^2_{gj}}{2\sum_i p_i^j q_i^j}$ where σ^2_{ej} and σ^2_{gj} , are variance components
203 estimated using Gibbs sampling and p_i^j and q_i^j are the allelic frequency at *i*-th locus for the
204 *j*-th breed. (*j* = Holstein, Brown Swiss or Simmental).
205 *MB-SNPBLUP.* Data of MB animals considered as an homogenous population were
206 analysed according to model [1]. A unique set of solutions for SNP effects were estimated
207 and then used to compute DGV of validation animals. The lambda ratio for the pooled
208 three-breed population was calculated as weighted average of SB variance components:

209
$$\lambda_{HBS} = \frac{\frac{n_H\hat{\sigma}^2_{eH} + n_B\hat{\sigma}^2_{eB} + n_S\hat{\sigma}^2_{eS}}{n_H + n_B + n_S}}{\frac{\hat{\sigma}^2_{gH}}{2\sum_i p_i^H q_i^H} + \frac{\hat{\sigma}^2_{gB}}{2\sum_i p_i^B q_i^B} + \frac{\hat{\sigma}^2_{gS}}{2\sum_i p_i^S q_i^S}}$$ where:

210 n_H, n_B and n_S are the population size for Holstein, Brown Swiss and Simmental respectively;
211 $\hat{\sigma}^2_{ej}$ and $\hat{\sigma}^2_{gj}$ are the estimated variance components (*j* = Holstein, Brown Swiss or
212 Simmental, respectively); p_i^j q_i^j are the allelic frequencies at *i*-th locus for the *j*-th breed.
213 Lambda ratios for the other MB combinations were calculated in the same way.
214 *MB-PC.* Effects of principal components were estimated with model [1] by replacing SNP
215 genotypes with PC scores in **Z.** Different lambda ratios were calculated for each MB
216 scenario following the approach of MB-SNPBLUP.

For each model, MME were solved by using a Gauss-Seidel iterative method. SNP or PC effects ($\hat{\mathbf{g}}$) were then used to calculate DGV of validation bulls as:

$$\mathbf{DGV} = \mu + \mathbf{Z}\hat{\mathbf{g}}$$

Assessment of model accuracy.

Pearson correlation coefficients between DGV and DRGP (r_{DGV}) scaled by the squared root of the mean DRGP reliability (REL_{DRGP}), were used to evaluate DGV accuracy ($r_{DGV} = r_{DRGP,\,DGV} \big/ \sqrt{REL_{DRGP}}$). The scaling was aimed at accounting for inaccuracy of the phenotypes used in the genomic evaluation (Hayes *et al.*, 2009a, Calus *et al.*, 2013). It does not have any effect on r_{DGV} when REL_{DRGP} is equal to one. Furthermore, the correlation between DGV and Pedigree Index (PI) was calculated (r_{PI}). Slope of the regression of DRGP on DGV was also calculated to evaluate the different models. Both r_{DGV} and b_{DGV} were calculated separately for each breed, for both SB and MB scenarios.

Results

Principal component analysis

The patterns of eigenvalues obtained for the different chromosomes in the SB and MB scenarios are reported in Figure 1. It is a useful tool for a visual detection of PC that met the eigenvalue >1 criterion. Principal components are extracted in order to maximize successively the amount of the original variance explained. Hence, the first component has the largest eigenvalue (i.e. the variance accounted for), the second PC the maximum after the first, and so on. Thus, the plot of eigenvalues is commonly characterized by a drop as the PC extraction proceeds. In the present study, such a drop was more

240 pronounced for the breeds with the smallest number of genotyped animals (i.e. Simmental
241 and Brown Swiss).
242
243 **Figure 1**
244 The variance accounted for by retained PCs varied from 0.85 (±0.01) in HBS to 0.92
245 (±0.01) in Brown Swiss scenario, corresponding to 7284 and 4029 PC, respectively. The
246 average number of PC retained per chromosome ranged from 149±42 (Brown Swiss) to
247 226±65 (HBS). The Simmental showed the largest number of PC in comparison to the
248 small size of its population (Table 3).
249
250 **Table 3**
251 Figure 2 reports individual PC scores for the first three principal components. Although
252 they were able to explain only about 9% of the original variance, the three breeds are
253 clearly separated. In particular, the first PC separates Holstein from the other two breeds,
254 whereas Brown Swiss and Simmental clustered in two different group along the second
255 PC. The third PC summarizes the interior variability of the largest group of bulls (Holstein).
256 **Figure 2**
257
258
259 *Genomic prediction accuracy*
260 *SB-SNPBLUP.* DGV accuracies for both SB and MB scenarios are reported in Table 4. In
261 the SB scenarios the accuracy varied across breeds and traits. The highest value was
262 observed in Holstein, the lowest in Brown Swiss. The accuracy of DGV was in most cases
263 higher than accuracy of pedigree index. However, in Brown Swiss r_{DGV} was lower than r_{PI}
264 for MY and PY (Table 4).

265 *SB-PC.* PCA reduced the number of predictors by 85% (±3%) on average. However, r_{DGV}
266 for Holstein decreased by about 5% when PC scores instead of SNP genotypes were used
267 as predictors. Conversely, the application of PCA did not affect r_{DGV} in the other two
268 breeds (Table 4).
269
270 *MB-SNPBLUP.* The combination of a multi breed reference population with the SNPBLUP
271 model did not affect the average r_{DGV} in comparison to the single breed scenario. If
272 compared to SB-SNPBLUP, the maximum r_{DGV} difference were +3% (HS) in Simmental
273 validation. With the exclusion of Holstein, the application of MB-SNPBLUP produced
274 similar r_{DGV} if compared to SB-PC.
275
276 *MB-PC.* In general, the use of a MB-PC slightly affected r_{DGV} compared to the other
277 models. In Holstein, an average r_{DGV} difference of +4% (vs SB-PC), -1% (vs SB-
278 SNPBLUP) and no difference (vs MB-SNPBLUP) were observed when HBS instead of
279 single breed was used as reference, respectively. Average accuracy did not change in
280 Simmental for MB-PC scenario, whereas slight differences of r_{DGV} were observed
281 compared to MB-SNPBLUP. Increases of 2% and 5% (vs SB-SNPBLUP) were observed
282 for Brown Swiss using HBS and HB reference population respectively. However, an
283 average decrease of 2% (vs SB-PC) and 4% (vs SB-SNPBLUP) was found using BS as
284 reference (Table 4). Looking at MB scenarios, most of the results are fairly comparable.
285 MB-PC average r_{DGV} difference spanning from -3% (BS) up to +4% (HB) if compared to
286 MB-SNPBLUP in Simmental and Brown Swiss validation set, respectively.
287

As far as DGV accuracy across traits is concerned, no clear pattern may be observed in the different MB scenarios (Table 4). The use of MB-PC was advantageous for Brown Swiss over SB for MY and PY. For instance, r_{DGV} of MY nearly doubled when Holstein were also present in the reference (HB +13% and +16% vs SB-PC and SB-SNPBLUP respectively). These gains were reduced (+7% SB-PC and +10% SB-SNPBLUP) when also Simmental was included in the reference (HBS scenario), whereas a drop in r_{DGV} was observed by combining Brown Swiss and Simmental (−4% SB-PC and -1% SB-SNPBLUP). A similar pattern can be observed for PY, with gain of reduced magnitudes. Conversely, a reduction of r_{DGV} was obtained for FY in all MB scenarios especially when Holstein bulls were in the reference population (−9%HB, −7% HBS, and −1% BS, respectively). Accuracy of DGV increased across different MB scenarios for yield traits in Holstein: up to 6%, 5% and 2% for MY, PY and FY, respectively (HBS reference).

Table 4

Pearson correlations between DGV for validation bulls calculated using MB-PC or SB-PC approaches are reported in Table 5. Across traits and MB reference population, the correlation ranged from 0.89 to 0.93, from 0.67 to 0.91 and from 0.88 to 0.98 for Holstein, Brown Swiss and Simmental, respectively. Very similar values for different validation set were observed across traits. Holstein did not show variation of correlations among different MB references and presented the highest value for FY (0.93). DGV calculated for Simmental using BS reference were highly correlated with DGV estimated using Simmental only for all the traits (>0.97). The correlation among SB and MB DGV of Brown Swiss were lower for the breed combinations HB and HBS (from 0.67 to 0.74 depending on the trait) in comparison to the breed combination BS (0.91).

312 **Table 5**

313 Table 6 reports the regression slopes of DGV on DRGP in SB and MB scenarios
314 using SNPBLUP or PC approaches. For SB-PC and MB-PC, the regression slopes were
315 fairly lower than 1 for all scenarios denoting a bias of prediction. No substantial changes
316 were observed for Holstein passing from SB to MB. In general, the bias of prediction was
317 higher both in Brown Swiss and Simmental when MB-PC genomic evaluations were
318 carried out, with a generalized reduction of the regression coefficients.

319 DGV estimates of the SNPBLUP models were biased as well, albeit that the
320 magnitude of the bias was smaller than for the PC models.

321 **Table 6**

322

323 **Discussion**

324 *Principal component analysis*

325 In the present work a multivariate SNP reduction method was tested both in single and
326 multi-breed populations and compared with the conventional approach of using SNP
327 genotypes as predictors.

328 The determination of the number of components to retain represents a crucial problem that
329 researcher must handle when using PCA. In fact, an incorrect choice may imply the
330 under-extraction of components, can lead to the loss of relevant information and it is likely
331 to introduce distortion in the solutions (Ledesma and Valera-Moro 2007). On the other
332 hand the extraction of a redundant number of PC may also be possible with less serious
333 consequences. In the current investigation, the number of retained PC was based on the
334 definition of a threshold for eigenvalues extracted from chromosome-wise SNP correlation
335 matrices. Cross validation procedures or Montecarlo simulations are often used to
336 establish the significant number of PC (Ledesma and Valera-Moro 2007). In genomic

337 selection framework, different approaches have been proposed. For instance, in
338 supervised PC Regression proposed by Long *et al.* (2011), a panel of SNP was
339 preselected according to associations with phenotypes and then PC were extracted. An
340 increase of genomic prediction accuracy was observed for PC extracted from the selected
341 SNP panel in comparison with the PCA carried out on the whole set of SNP (Long *et al.*,
342 2011). However, in this approach the number of retained PC may change across
343 phenotypes. Whereas, The MINEIGEN criterion was adopted in the present work for
344 identifying the optimum amount of variance accounted for PC in datasets of different size
345 and for any traits. Despite some criticism on the MINEIGEN criterion, it is still valid for
346 decomposition of correlation matrix with unities at the diagonal elements (Ledesma and
347 Valera-Moro 2007).
348
349 The retained PC were able to explain comparable amounts of variance (~90%) in
350 the three breeds for the SB scenario. Despite that, a higher number of PC were found for
351 Simmental. This feature was already observed in our previous work (Pintus *et al.*, 2012)
352 and it can be explained by differences in the genetic structure of this population (e.g.
353 Linkage Disequilibrium pattern, see later in the discussion), or by an overestimation of the
354 significant number of PC able to best explain original correlation among SNP variables
355 (Ledesma and Valero-Mora, 2007). Although the number of PC retained was higher than
356 previous reports (Long *et al.*, 2011, Pintus, 2012) a considerable reduction of the predictor
357 dimensionality was achieved though.
358
359 *Genomic prediction accuracy*
360 *SB approach.* The average DGV accuracy for SB-SNPBLUP model in Holstein, Brown
361 Swiss and Simmental reflects somehow the difference in the size of the reference

362 populations, as previously observed in the same (Pintus *et al.*, 2012, Pintus *et al.*, 2013) or
363 other Holstein populations of similar size (VanRaden *et al.*, 2009).
364 The application of SB-PC in Holstein resulted in a large reduction of predictor
365 dimensionality, with some negative effects on predictive ability. The reduction in r_{DGV} were
366 systematic, and probably related to the number of retained PC. A substantial equivalence
367 among PC and other methods was highlighted, in our previous work, when a larger
368 number of PC was extracted (15609 *vs* 4908 used in the present paper) (Pintus *et al.*,
369 2013). Conversely, no substantial changes (or slight improvement) in r_{DGV} were observed
370 in Simmental and Brown Swiss in comparison to SB-SNPBLUP. For Simmental, the r_{DGV}
371 of PY was lower than values obtained by Gredler *et al.*, (2009), Gredler *et al.*, (2010)
372 using a Partial Least Squares Regression approach. However in both cases the reference
373 population size was larger than in the present work (1091 and 2477 bulls, respectively).
374 DGV accuracies for Brown Swiss were consistent with our other previous work, but lower
375 than those reported in literature. For instance the r_{DGV} for PY was 0.16 in comparison to
376 0.32 (Olson *et al.*, 2012), 0.55 (Olson *et al.*, 2011) and 0.60 (Jorjani *et al.*, 2012) using
377 reference population of 506 (US), 1056 (US) and 4800 (InterGenomics) Brown Swiss bulls
378 respectively. This fact clearly denotes the effect of population size on DGV accuracy.
379 *MB approach.* The application of MB slightly improved average r_{DGV} of yield traits in
380 comparison to SB-PC. Across multibreed scenarios, MB-SNPBLUP and MB-PC performed
381 similarly. The r_{DGV} for Holstein were lower than those reported by Hayes *et al.*, (2009b)
382 even if they used less animals in the reference. In a work of Pryce *et al.*, (2011), after a
383 further enlargement of the previous MB reference population (including Holstein,
384 Simmental and Jersey) no substantial changes in the r_{DGV} were recorded for milk
385 production traits.

386 Looking at specific traits, some interesting results came up, even if without a clear
387 and constant pattern across MB scenarios. Difference among traits are probably due to
388 the sampling effect due to the reduced size of the population involved in the present work.
389 However, interesting pattern in r_{DGV} can be observed among traits. The highest gain in
390 r_{DGV} for MY and PY was observed by pooling Holstein and Brown Swiss population
391 together. A partial decrease was observed when Simmental was added to the dataset
392 (HBS), whereas the combination BS gave the worst results. Brown Swiss and Simmental
393 together presented the largest difference at LD level (Figure 3), and this could explain the
394 reduced accuracy of milk traits from their combination.
395 Presented results are in agreement with reports on Nordic Red Cattle (Brondum *et*
396 *al.*, 2011). In particular MB genomic evaluation produced gain of 7% and 9% for MY
397 (+10% for PY) in Swedish and Finnish validation populations respectively. Adding a third
398 breed (Danish Red) sometime was beneficial for the other two breed, whereas just slight
399 gain in accuracy were recorded for Danish itself, across different traits. Their results
400 probably rely on similar LD among breeds (0.20) (Brondum *et al.*, 2011) and particularly on
401 reduced genetic distances between Swedish and Finnish cattle. In fact, these two breeds
402 are of the Ayrshire type, while the Danish Red has some old influence from Brown Swiss
403 and Holstein (Brondum *et al.*, 2011). A similar pattern may be observed in our dataset for 2
404 traits under control of many genes such as PY and MY. Indeed, Brown Swiss and Holstein
405 have similar Linkage Disequilibrium patterns (Figure 3) and probably this similarity makes
406 possible to pick up QTL effects across breeds using PCA. However, this conclusion is not
407 supported by the literature. For instance, in US Brown Swiss just a slight increase of
408 accuracy was achieved by adding Holstein in the reference population. This fact was
409 probably due to small contribution (less than 10%) of Brown Swiss to the whole MB

410 population (Olson *et al.*, 2012) in comparison to our dataset (30% and 20% of Brown
411 Swiss in HB and HBS respectively).

412 **Figure 3**

413 If MY and PY showed an increase of accuracy in MB scenarios, opposite behavior
414 was observed for FY, specially for MB-PC approach. The genetic background of FY may
415 explain these results. It is known that a polymorphism in DGAT1 gene (BTA 14) explains
416 >40% of genetic variance of FY, whereas the genetic background of MY and PY is
417 markedly polygenic. Despite DGAT1 polymorphism is not included in the 54K panel, SNP
418 markers in LD with this gene can capture part of its genetic variance. DGAT1 is
419 segregating in the Italian Holstein population, but not in Italian Brown Swiss (Scotti *et al.*,
420 2010). Hence, PC effects mighty be biased by the fact that in Italian Brown Swiss and
421 Italian Simmental one of the allele is fixed. This hypothesis need to be verified but the
422 comparison of PC effects on BTA 14 both in SB and MB scenarios might have led to such
423 conclusion. (Figure 4). In fact, no large effects were observed on BTA14 for Italian Brown
424 Swiss and Italian Simmental. Conversely, in Holstein a big PC signal was found on BTA14
425 as well as in any MB scenario including Holstein.

426 **Figure 4**

427 In general, prediction biases were observed in our model. In all cases regression slopes of
428 DRGP on DGV were lower than one, indicating inflation of variance for all prediction
429 methods. An optimal prediction would led to regression slope of 1, in the present work the
430 DGV estimates are inflated in both MB and SB scenarios, even if BLUP estimates
431 presented b_{DGV} coefficients slightly higher than PC scenarios. A clear pattern across traits
432 and scenarios hardly can be identified, likewise other MB papers (Brondum *et al.*, 2011).
433 Moreover, prediction bias increased for MB in comparison to either SB analysis in the
434 present paper or other work involving the same populations (Pintus *et al.*, 2012, Pintus *et*

435 *al.*, 2013), and this could be also due to the increase of the dimensionality of predictors.
436 Another possible explanation is related to the expected value of slopes: b_{DGV} is 1 only if the
437 genotyped animals are a representative sample of the animals population in the
438 corresponding age classes (Mäntysaari *et al.*, 2011; Patry *et al.*, 2013). For Simmental and
439 Brown almost all the available bulls were genotyped, whereas a bias can be introduced by
440 selecting the Holstein bulls from a larger population. Probably, in MB reference population
441 (with a higher proportion of Holstein) an expected values for b_{DGV} different from one could
442 be hypothesized, depending selective genotyping of bulls. Biases in genomic predictions
443 can also be due to the multi-step genomic selection procedure in population under
444 selection. The application of prediction equations developed in training population using
445 pseudo-phenotypes as observations (DRGP) was claimed to introduce bias in the DGV
446 (Vitezica *et al.*, 2011). Inflation of DGV variance were also observed in other works that
447 use multivariate regression methods for genomic prediction. For instance, Solberg *et al.*,
448 (2009) found that the b_{DGV} decreases as the number of latent variables used grew. In
449 multivariate context, this problem can be overcome by cross validation to identify the
450 number of PC that provide unbiased estimate of DGV (Solberg *et al.*, 2009).
451 *General discussion*
452 The summary of DGV accuracy as function of the reference population size, obtained in
453 the present work, together with some of the results retrieved from recent literature is
454 presented in Table 7. The increase in population size pooling together multiple breed
455 populations gave rise just to slight increase in DGV accuracy according to most of reported
456 results. Figures in Table 7 might suggest that MB approach works better when breeds are
457 not too genetically distant, especially for some of the Nordic Red Cattle. For reference
458 population of reduced size, an apparent overestimations of DGV accuracy was observed

459 for some breeds, whereas there are other cases of underestimation as Brown Swiss in our
460 data. Actually, a possible explanation for this apparent overestimation can be found in the
461 different strategy for the calculation of GEBV reliability implemented in diverse genomic
462 evaluation softwares.

463 **Table 7**

464 In order to try to explain these results of accuracy the within breed LD level was
465 investigated. The patterns of LD in Simmental and Holstein populations are in agreement
466 to the finding of Pryce *et al.*, (2011) in Australian Holstein and German Fleckvieh. The LD
467 values at the average marker distance in the 54K panel (about 67 kbp) were similar
468 between Brown Swiss and Holstein (0.19) and slightly lower in Simmental (0.15). For the
469 latter a lower LD persistency was also observed, with a sharp drop of LD over short
470 distance in comparison to Holstein and Brown Swiss. Although Simmental had similar
471 number of genotyped bulls compared to Brown Swiss, its effective population size (N_e) is
472 greater. That was expected to have a negative effect on the accuracy of genomic
473 prediction of Simmental but did not. A possible explanation is that a fair number of Brown
474 Swiss bulls (~1/4) were born before 1980 (and the oldest bull dates 1960) in contrast to
475 the Simmental and Holstein reference population whose bulls were more closer to each
476 other (Pintus *et al.*, 2012, Pintus *et al.*, 2013). Another possible explanation could be found
477 in the influence of relatedness between reference and validation populations (96 and 70
478 father son pairs were included in the Brown Swiss and Simmental population, respectively)
479 as also hypothesized by Habier *et al.*, (2010) and Pszczola *et al.*, (2012).

480

481 *Conclusions*

482 Results of the present study showed a slight average increase of DGV accuracy in the
483 multi-breed approach compared to the single breed, although differences have been

484 observed between breeds. In particular, r_{DGV} seemed to be quite in agreement to the
485 theoretical expectation for Holstein, whereas Simmental did not exhibit gains in accuracy
486 using an MB reference population. Brown Swiss showed an increase of DGV accuracy in
487 MB scenarios for PY and MY and a decrease for FY. Differences in the LD structure of the
488 three breeds and in their sample size may explain at least partially these results. Within
489 the MB approaches, basically no clear differences in DGV accuracy were observed
490 between the use of SNP genotypes or principal component scores as predictors.

491

492 **Acknowledgement**

493 Research funded by the Italian Ministry of Agriculture (grant INNOVAGEN) and by the
494 Fondazione CARIPLO (grant PROZOO). The research leading to these results has
495 received funding from the European Union's Seventh Framework Programme (FP7/2007-
496 2013) under grant agreement n° 289592 – Gene2Farm. Authors wish also to acknowledge
497 Italian Holstein (ANAFI), Brown Swiss (ANARB) and Simmental (ANAPRI) association for
498 providing phenotypic data.

499

500

References

Brondum RF, Rius-Vilarrasa E, Stranden I, Su G, Guldbrandtsen B, Fikse WF and Lund MS 2011. Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. Journal of Dairy Science 94, 4700-4707.

Calus MPL, de Haas Y and Veerkamp RF 2013. Combining cow and bull reference populations to increase accuracy of genomic prediction and genome-wide association studies. Journal of Dairy Science 96, 6703-6715.

Daetwyler HD, Kemper KE, van der Werf JHJ and Hayes BJ 2012. Components of the accuracy of genomic prediction in a multi-breed sheep population. Journal of Animal Science 90, 3375-3384.

de Roos APW, Hayes BJ and Goddard ME 2009. Reliability of Genomic Predictions Across Multiple Populations. Genetics 183, 1545-1553.

Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, Mason BA and Goddard ME 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science 95, 4114-4129.

Goddard ME and Hayes BJ 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics 10, 381-391.

Gredler B, Nirea KG, Solberg TR, Egger-Danner C, Meuwissen T and Sölkner J 2009. A comparison of methods for genomic selection in austrian dual purpose simmental cattle. Proceeding of the 18th Conference Advancement of Animal Breeding and Genetics, 28th Semptember 2009, Barossa Valley, South Australia pp. 568–571.

Gredler B, Schwarzenbacher H, Egger-Danner C, Fuerst C, Emmerling R and Sölkner J 2010. Accuracy of genomic selection in dual purpose Fleckvieh cattle using three types of methods and phenotypes. In Proceeding of 9th World Congress of genetics applied to livestock production,1st - 6th August 2010, Leipzig, Germany, Article n. 0907.

Habier D, Tetens J, Seefried FR, Lichtner P and Thaller G 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution 42, 5

Harris BL, Creagh FE, Winkelman AM and Johnson DL 2011. Expereinces with the Illumina High Density Bovine BeadChip. Interbul bulletin 44, 3-7.

Hayes BJ, Bowman PJ, Chamberlain AJ and Goddard ME 2009a. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science 92, 433-443.

Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K and Goddard ME 2009b. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution 41, 51

Jorjani H, J. Jakobsen, E. Hjerpe, V. Palucci and J. Dürr. 2012. Status of Genomic Evaluation in the Brown Swiss Populations. Interbull Bulletin 46, 46-54.

Kaiser HF 1960. The Application of Electronic Computers to Factor Analysis. Educational and Psychological Measurement 20, 141-151.

Karoui S, Carabano MJ, Diaz C and Legarra A 2012. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. Genetics Selection Evolution 44, 39.

Kizilkaya K, Fernando RL and Garrick DJ 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. Journal of Animal Science 88, 544-551.

Legarra A, Ricard A and Filangi O 2012. GS3 Manual User (genomic selection, Gibbs sampling Gauss Seidel). Retrieved on 12 December 2013 from http://snp.toulouse.inra.fr/~alegarra/manualgs3_last.pdf

549 Ledesma RD and Valero-Mora P 2007. Determinig the number of factors to retain in EFA: an easy-
550 to-use computer program for carryning out Parallel Analysis. Practical assessment, research
551 & evaluation 12.
552 Long N, Gianola D, Rosa GJM and Weigel KA 2011. Dimension reduction and variable selection
553 for genomic selection: application to predicting milk yield in Holsteins. Journal of Animal
554 Breeding and Genetics 128, 247-257.
555 Lund MS, Roos AP, Vries AG, Druet T, Ducrocq V, Fritz S, Guillaume F, Guldbrandtsen B, Liu Z,
556 Reents R, Schrooten C, Seefried F and Su G 2011. A common reference population from
557 four European Holstein populations increases reliability of genomic predictions. Genet Sel
558 Evol 43, 43.
559 Macciotta NPP, Gaspa G, Steri R, Nicolazzi EL, Dimauro C, Pieramati C and Cappio-Borlino A
560 2010. Using eigenvalues as variance priors in the prediction of genomic breeding values by
561 principal component analysis. Journal of Dairy Science 93, 2765-2774.
562 Makgahlela ML, Mantysaari EA, Stranden I, Koivula M, Nielsen US, Sillanpaa MJ and Juga J 2013.
563 Across breed multi-trait random regression genomic predictions in the Nordic Red dairy
564 cattle. Journal of Animal Breeding and Genetics 130, 10-19.
565 Mäntysaari E, Liu Z and VanRaden P 2011. Interbull Validation Test for Genomic Evaluations.
566 Interbull Bulletin 41, 17-21.
567 Meuwissen THE, Hayes BJ and Goddard ME 2001. Prediction of total genetic value using
568 genome-wide dense marker maps. Genetics 157, 1819-1829.
569 Olson KM, VanRaden PM and Tooker ME 2012. Multibreed genomic evaluations using purebred
570 Holsteins, Jerseys, and Brown Swiss. Journal of Dairy Science 95, 5378-5383.
571 Olson KM, VanRaden PM, Tooker ME and Cooper TA 2011. Differences among methods to
572 validate genomic evaluations for dairy cattle. Journal of Dairy Science 94, 2613-2620.
573 Patterson N, Price AL and Reich D 2006. Population Structure and Eigenanalysis. Plos Genetics 2,
574 e190.
575 Pintus MA, Gaspa G, Nicolazzi EL, Vicario D, Rossoni A, Ajmone-Marsan P, Nardone A, Dimauro
576 C and Macciotta NP 2012. Prediction of genomic breeding values for dairy traits in Italian
577 Brown and Simmental bulls using a principal component approach. Journal of Dairy Science
578 95, 3390-3400.
579 Pintus MA, Nicolazzi EL, Van Kaam JBCHM, Biffani S, Stella A, Gaspa G, Dimauro C and
580 Macciotta NPP 2013. Use of different statistical models to predict direct genomic values for
581 productive and functional traits in Italian Holsteins. Journal of Animal Breeding and Genetics
582 130, 32-40.
583 Pryce JE, Gredler B, Bolormaa S, Bowman PJ, Egger-Danner C, Fuerst C, Emmerling R, Solkner
584 J, Goddard ME and Hayes BJ 2011. Short communication: Genomic selection using a multi-
585 breed, across-country reference population. Journal of Dairy Science 94, 2625-2630.
586 Pszczola M, Strabel T, Mulder HA and Calus MPL 2012. Reliability of direct genomic values for
587 animals with different relationships within and to the reference population. Journal of Dairy
588 Science 95, 389-400.
589 Scotti E, Fontanesi L, Schiavini F, La Mattina V, Bagnato A and Russo V 2010. DGAT1 p.K232A
590 polymorphism in dairy and dual purpose Italian cattle breeds. Italian Journal of Animal
591 Science 9, 79-82.
592 Solberg TR, Sonesson AK, Woolliams JA and Meuwissen THE 2009. Reducing dimensionality for
593 prediction of genome-wide breeding values. Genetics Selection Evolution 41, 29
594 VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF and
595 Schenkel FS 2009. Invited review: Reliability of genomic predictions for North American
596 Holstein bulls. Journal of Dairy Science 92, 16-24.
597 VanRaden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME, Cole JB, Sonstegard TS, Connor
598 EE, Winters M, van Kaam JBCHM, Valentini A, Van Doormaal BJ, Faust MA and Doak GA

599 2013. Genomic imputation and evaluation using high-density Holstein genotypes. Journal of
600 Dairy Science 96, 668-678.
601

602 **Tables**

603

604 **Table 1**. Composition of Reference and Validation populations used for DGV estimation in
605 Italian Holstein (Hol), Italian Simmental (Sim) and Italian Brown Swiss (Brw) single breed
606 (SB) cattle breed or Multiple Breed (MB) population. Number of bulls left after data editing
607 and cut-off year of birth used to define reference and validation population were reported
608 both for SB or MB population.

Reference Population	Validation Population	Birth years	No Bull[1]	Bulls Used[2]	Ref Year ≤ 2000	Val[3] Year > 2000
Single Breed (SB)						
Hol	Hol	1979-2007	2093	2058	1424	634
Sim	Sim	1972-2006	717	551	380	171
Brw	Brw	1960-2004	749	634	493	141
Multi Breed (MB)						
HBS	(Hol+Brw+Sim)	-	3559	3245	2299	634+171+141
HS	(Hol+Sim)	-	2810	2610	1805	634+171
HB	(Hol+Brw)	-	2842	2692	1917	634+141
BS	(Brw+Sim)	-	1466	1185	873	171+141.

609 [1] Bulls used in Principal component analysis

610 [2] Bulls used for genomic evaluation: differences in the number of bulls are due to missing phenotypes.

611 [3] Validation bulls of MB dataset were the same as single breed analysis:

612

613

614 **Table 2**. Number of SNP retained after data editing and related causes of elimination
615 (MAF = minor allele frequency, HW=Hardy-Weinberg) for Holstein (Hol), Simmental (Sim)
616 and Brown Swiss (Brw) considered both as separate (Hol, Brw, Sim) or pooled population
617 (HBS, BS, HB, HS).

618

	SB			MB			
Cause of elimination	Hol	Sim	Brw	HBS	HS	HB	BS
Monomorphic	5481	5416	6282	4337	4553	4703	4931
MAF < 5%	5521	6376	8319	4786	4973	4998	6417
Callrate <97.5%	1633	1314	818	1489	1614	1499	1025
No heterozygous	166	107	176	105	104	173	106
Not HW equilibrium	197	161	119	0	0	0	0
Mendelian conflict	64	185	84	263	254	207	246
SNP Discarded	*13100*	*13559*	*15819*	*10980*	*11498*	*11581*	*12725*
SNP Used	*39240*	*38781*	*36521*	*41360*	*40842*	*40759*	*39615*

619

620 **Table 3.** Average variance explained by PC (%) for SB and MB datasets, average number
621 of PC by chromosome and total number of PC used. Number of rows and columns of
622 chromosome-wise SNP correlation matrices.

Dataset[1]	Variance explained (%)[2]	Average number of PC[3] ± sd	PC used	No. row (n bulls)
Brw	92	149 ± 42	4029	749
Sim	91	218 ± 57	6402	717
Hol	90	160 ± 43	4908	2093
HB	88	188 ± 53	5840	2482
HS	87	211 ± 60	7099	2810
BS	86	212 ± 43	6477	1466
HBS	85	226 ± 65	7284	3559

623 [1] Brw=Brown Swiss, Sim=Simmental, Hol=Holstein, HB=(Hol+Brw), HS=(Hol+Sim), BS=(Brw+Sim),
624 HBS=(Hol+Brw+Sim).
625 [2] Variance explained by all PCs which eigenvalues was >1 averaged by 29 chromosome (standard deviation
626 1%)
627 [3] These values represent the average across 29 chromosomes.
628

Table 4. Realized Pedigree Index accuracy (r_{PI}) for Milk, Fat and protein Yield. DGV accuracy (r_{DGV}) for single breed (SB) approach using the whole set of markers (SB-SNPBLUP) or principal component analysis (SB-PC). Multiple breed DGV accuracy using SNPBLUP (MB-SNPBLUP) or PCA approaches (MB-PC) for different combination of reference population.

Single breed (SB)	(r_{PI})			SNPBLUP (r_{DGV})			PC (r_{DGV})		
Validation[1]	Hol	Brw	Sim	Hol	Brw	Sim	Hol	Brw	Sim
Milk Yield	0.45	0.21	0.34	0.45	0.13	0.38	0.39	0.16	0.38
Fat Yield	0.34	0.23	0.33	0.45	0.28	0.32	0.42	0.27	0.35
Protein Yield	0.40	0.20	0.34	0.41	0.14	0.36	0.36	0.16	0.36
Average	0.40	0.21	0.34	0.44	0.18	0.35	0.39	0.20	0.36
Sd	0.06	0.02	0.01	0.02	0.08	0.03	0.03	0.06	0.02

Multiple Breed (MB)	SNPBLUP (r_{DGV})								
Reference[2]	HBS			HB		HS		BS	
Validation[1]	Hol	Brw	Sim	Hol	Brw	Hol	Sim	Brw	Sim
Milk Yield	0.45	0.17	0.38	0.45	0.18	0.45	0.39	0.13	0.38
Fat Yield	0.44	0.26	0.34	0.44	0.24	0.44	0.37	0.29	0.31
Protein Yield	0.42	0.16	0.37	0.41	0.16	0.41	0.39	0.14	0.36
Average	0.44	0.19	0.36	0.44	0.19	0.44	0.38	0.19	0.35
Sd	0.02	0.05	0.02	0.02	0.05	0.02	0.01	0.09	0.04
	PC (r_{DGV})								
Reference[2]	HBS			HB		HS		BS	
Validation[1]	Hol	Brw	Sim	Hol	Brw	Hol	Sim	Brw	Sim
Milk Yield	0.45	0.23	0.37	0.43	0.29	0.44	0.36	0.12	0.38
Fat Yield	0.44	0.20	0.34	0.44	0.18	0.44	0.35	0.26	0.33
Protein Yield	0.41	0.17	0.36	0.39	0.21	0.40	0.37	0.11	0.36
Average	0.43	0.20	0.36	0.42	0.23	0.43	0.36	0.16	0.36
Sd	0.02	0.03	0.02	0.03	0.06	0.02	0.01	0.08	0.03

[1] Hol=Holstein (n=634) ; Brw=Brown Swiss (n=141), Sim=Simmental (n=171)

635 [2] HBS = Hol+Brw+Sim (n=2299); HB =Hol+Brw (n=1805); HS=Hol+Sim HS (n=1917); BS=Brw+Sim

636 (n=873)

637 **Table 5.** Pearson Correlation among DGV calculated for yield trait in validation bulls using

638 single breed (SB) and Multiple breed (MB) reference population.

		MB Reference[1]			
Trait	Validation[2]	HBS	HB	HS	BS
Milk yield	Hol	0.89	0.89	0.89	*
	Brw	0.70	0.67	*	0.91
	Sim	0.88	*	0.89	0.98
Fat Yield	Hol	0.93	0.93	0.93	*
	Brw	0.67	0.68	*	0.91
	Sim	0.84	*	0.87	0.97
Protein Yield	Hol	0.92	0.92	0.91	*
	Brw	0.74	0.74	*	0.91
	Sim	0.89	*	0.89	0.98

639 [1] HBS = Hol+Brw+Sim (n=2299); HB =Hol+Brw (n=1805); HS=Hol+Sim HS (n=1917); BS=Brw+Sim (n=873)

640 [2] Hol=Holstein (n=634) ; Brw=Brown Swiss (n=141), Sim=Simmental (n=171)

641

642 **Table 6.** Bias of prediction measured by b(DRGP,DGV) for single breed (SB) and multiple breed (MB) approach for yield traits using

643 PC or SNPBLUP methods.

PC	SB			MB								
Reference[1]	Hol	Brw	Sim	HBS			HB		HS		BS	
Validation[2]	Hol	Brw	Sim	Hol	Brw	Sim	Hol	Brw	Hol	Sim	Brw	Sim
Milk Yield	0.40	0.20	0.78	0.40	0.24	0.39	0.45	0.33	0.54	0.53	0.18	0.66
Fat Yield	0.49	0.38	0.67	0.47	0.19	0.42	0.50	0.20	0.49	0.45	0.32	0.60
Protein Yield	0.38	0.21	0.71	0.36	0.17	0.42	0.39	0.23	0.45	0.50	0.13	0.61
SNPBLUP	SB			MB								
Validation[2]	Hol	Brw	Sim	Hol	Brw	Sim	Hol	Brw	Hol	Sim	Brw	Sim
Milk Yield	0.65	0.20	0.72	0.64	0.26	0.68	0.63	0.28	0.65	0.71	0.17	0.71
Fat Yield	0.76	0.46	0.74	0.72	0.42	0.63	0.63	0.41	0.75	0.70	0.43	0.59
Protein Yield	0.54	0.22	0.78	0.54	0.23	0.62	0.53	0.24	0.55	0.66	0.21	0.65

644 [1] HBS = Hol+Brw+Sim (n=2299); HB =Hol+Brw (n=1805); HS=Hol+Sim HS (n=1917); BS=Brw+Sim (n=873)

645 [2] Hol=Holstein (n=634) ; Brw=Brown Swiss (n=141), Sim=Simmental (n=171)

646

647

Table 7. Average Genomic Selection accuracy[1] across yield traits as function of the size of the reference population. Data for DGV accuracy for milk, protein and fat yield were averaged from recent literature on multi-breed Genomic Selection.

				VALIDATION				
TYPE	BREED[2]	TRAINING[3]	N[4]	DK	SWE	FIN	ALL	REFERENCE[5]
SingleBreed	NRC	DK	778	0.47	0.10	0.13		Brondum et al., (2011)
		SWE	1395	0.12	0.35	0.42		
		FIN	1562	0.10	0.38	0.45		
MutliBreed		SWE+FIN	2957		0.47	0.55	0.52	
		DK+SWE+FIN	3735	0.49	0.50	0.49	0.53	
		SWE+FAY+OTH	3300				0.58	Makgahlela et al. (2012)
		SWE+FAY+OTH	3300				0.60	
				VALIDATION				
				BRW	HOL	JER	SIM	
SingleBreed	BRW	IT BRW	493	0.20				¶
		US BRW	506	0.32				Olson et al., (2012)
	HOL	AU HOL	755		0.43			Pryce et al., (2011)
		AU HOL*	781		0.51			Hayes et al., (2009b)
		IT HOL	1424		0.39			¶
		US HOL	5331		0.70			Olson et al., (2012)

	JER	AU JER	243			0.52		Hayes et al., (2009b)
		US JER	1361			0.71		Olson et al., (2012)
	SIM	IT SIM	380				0.36	¶
		GER SIM	1247				0.41	Pryce et al., (2011)
MutliBreed		IT BRW+IT SIM	873	0.16				¶
		IT HOL+IT BRW	1917	0.23	0.42			¶
		IT HOL+IT BRW+IT SIM	2299	0.20	0.43		0.36	¶
		US HOL+US JER+US BRW	7198	0.36	0.69	0.70		Olson et al., (2012)
		AU HOL+AU JER	1024		0.51	0.50		Hayes et al., (2009b)
		AU HOL+AU JER*	1141		0.41			Pryce et al., (2011)
		IT HOL+IT SIM	1805		0.43		0.36	¶
		AU HOL+GER SIM	2002		0.41		0.31	Pryce et al., (2011)
		AU HOL+GER SIM+AU_JER	2388		0.42		0.31	Pryce et al., (2011)
		FR HOL+NOR+MON	4896		0.64		0.52	Karoui et al., (2012)

[1] DGV accuracy were expressed as simple correlation. Squared correlation from literature were converted using the square root of the published accuracy values.

[2] NRC Nordic red Cattle, BRW Brown Swiss, HOL Holstein, JER Jersey, SIM Simmental or Fleckvieh

[3] Reference populations used in within or across breed genomic prediction. Danish (DK), Finnish (FIN) and Swedish Red (SWE) dairy cattle, Finnish Ayrshire (FAY), other breeds (OTH). AUSTRALIAN DAIRY: Australian Holstein (AU HOL), Australian Jersey (AU JER) Austrian & German Fleckvieh (GER SIM). FRENCH DAIRY: French Holstein (FR HOL), Monbeliarde (MON), Normande (NOR). US DAIRY: US Holstein (US HOL), US Jersey (US JER) and Brown Swiss (US BRW). ITALIAN DAIRY: Italian Holstein (IT HOL), Italian Simmental (IT SIM), Italian Brown Swiss (IT BRW).

[4] Number of animals of different reference populations used in within or across breed genomic prediction.

[5] References of the corresponding figures. ¶ refers to the results presented in the current papers applying PC Multibreed approach.

660
661
662

663 **Figure Captions**

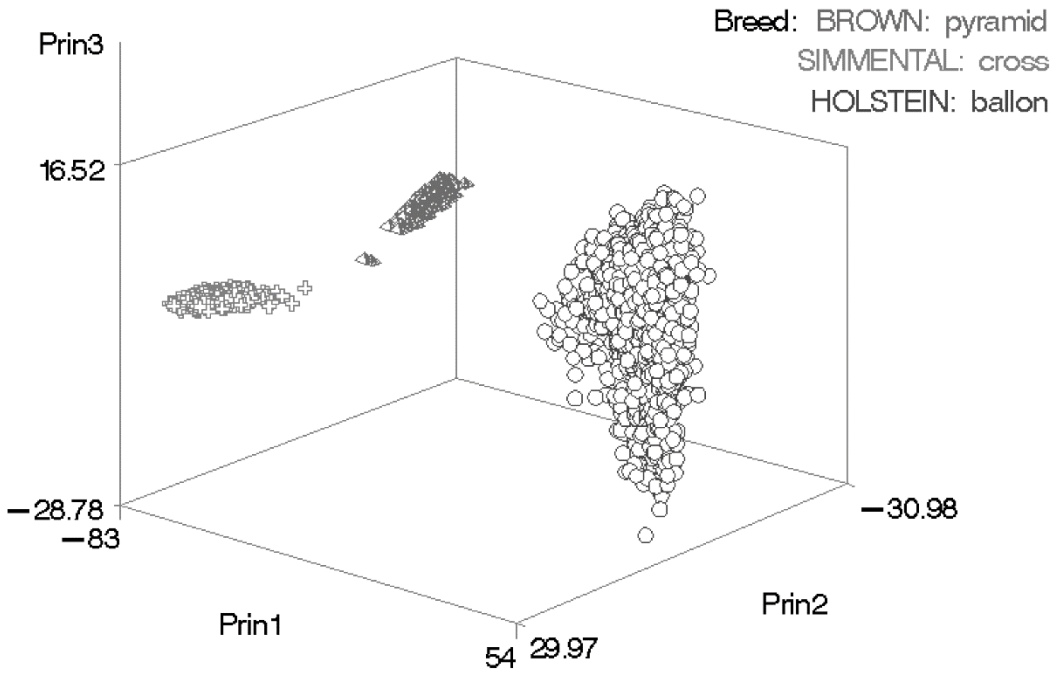
664 **Figure 1.** Pattern of eigenvalues as function of number of PC extracted: each line
665 represents the eigenvalues on logarithmic scale for each of the 29 chromosome analyzed
666 for Holstein (a), Brown (b), Simmental (c) and their combination (d).

667 **Figure 2.** Plot of the individual scores that animals belonging to different breeds obtained
668 on first three Principal Components (PC). (Variance explained by PC1=5.1%, PC2=2%,
669 PC3=1.6%).

670 **Figure 3.** Pattern of Linkage Disequilibrium (LD) within 1,000 kbp of distance among all
671 pairs of marker for Holstein (Hol), Brown Swiss (Brw) and Simmental (Sim), values
672 reported are the average r^2 across 29 chromosome.

673 **Figure 4.** Boxplots of PC or SNP effect estimates for fat yield in BTA14 in single breed
674 (Hol, Sim, Brw) or Multiple Breed reference population (HBS, HB, HS and BS).

675



676

677 Figure 1

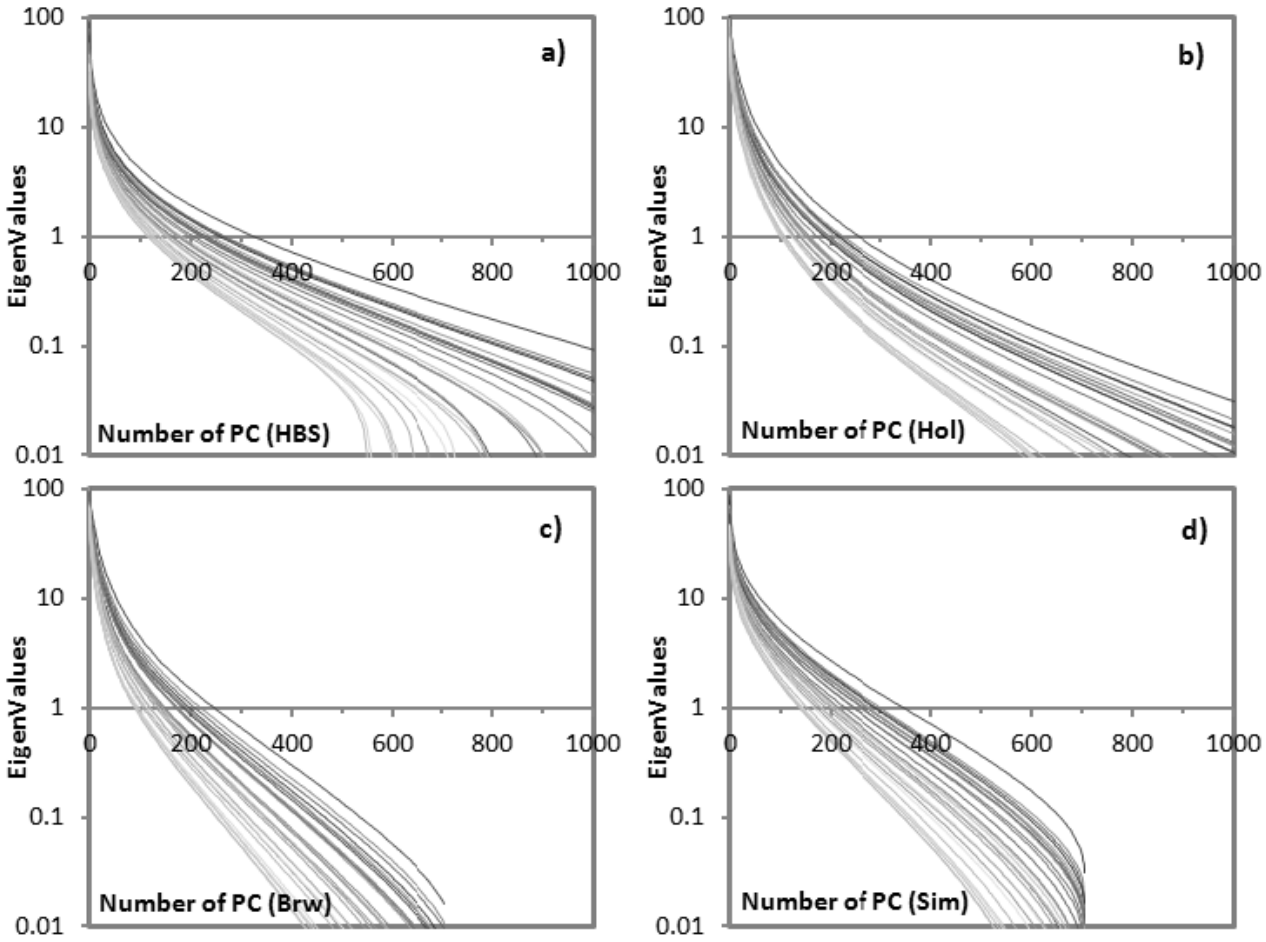
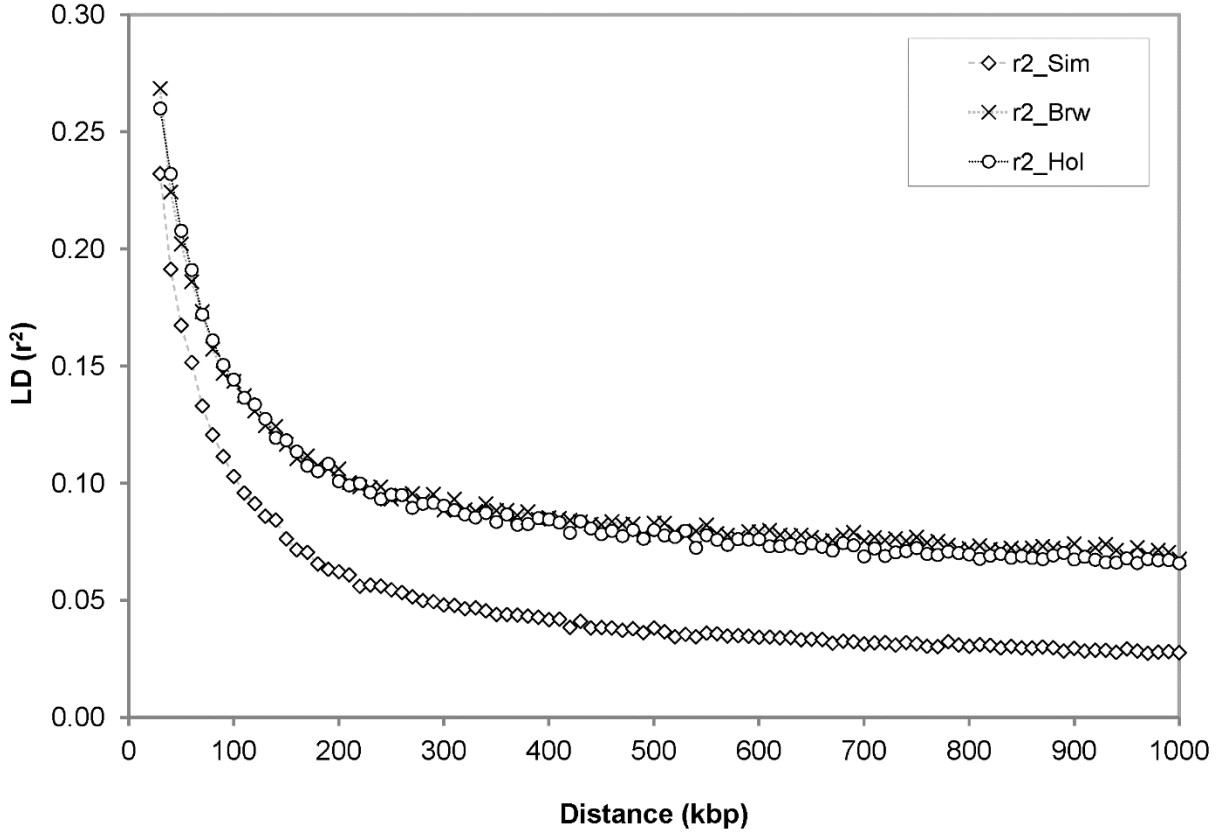


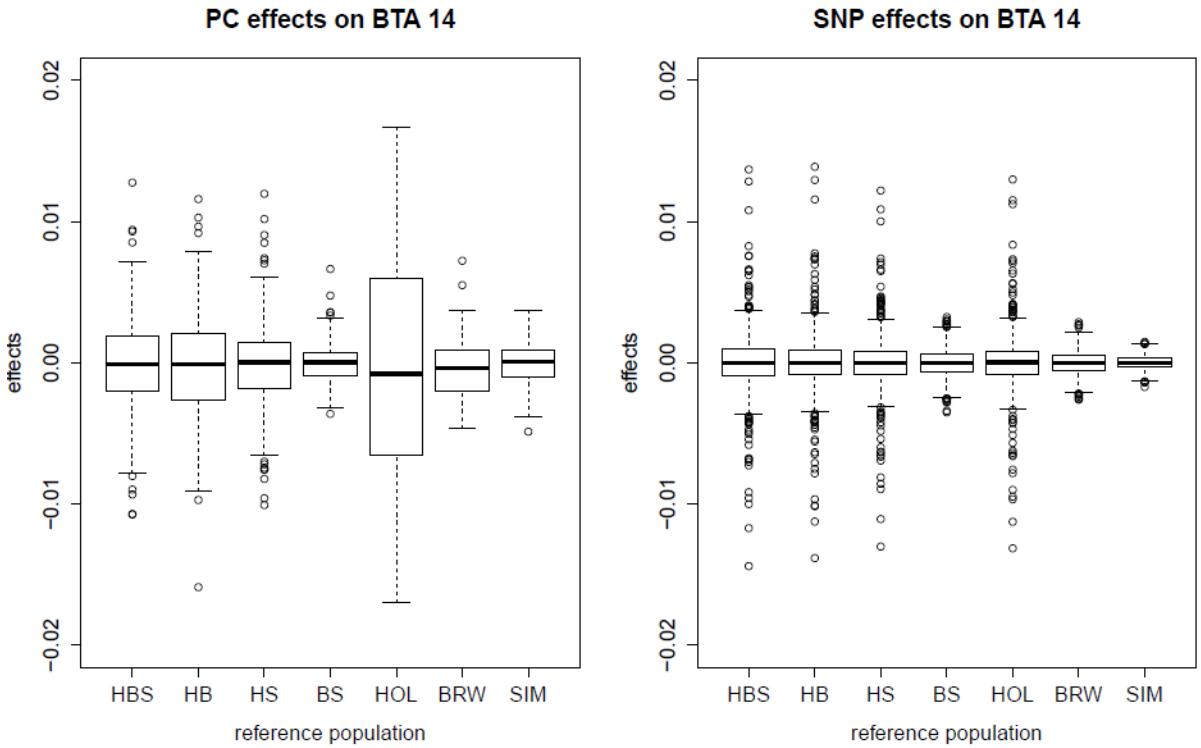
Figure 2



680

681 Figure 3

682



683

684 Figure 4